AvatarPose: Avatar-guided 3D Pose Estimation of Close Human Interaction from Sparse Multi-view Videos

Anonymous ECCV 2024 Submission

Paper ID #6979

Fig. 1: We propose AvatarPose, a method to estimate the 3D poses and shapes of multiple closely interacting people from multi-view videos. To this end, we first reconstruct the avatar of each individual and leverage the learned personalized avatars as priors to refine poses via color and silhouette rendering loss. We alternate between avatar refinement and pose optimization to obtain the final pose estimation.

Abstract. Despite progress in human motion capture, existing multi-view methods often face challenges in estimating the 3D pose and shape of multiple closely interacting people. This difficulty arises from reliance on accurate 2D joint estimations, which are hard to obtain due to occlusions and body contact when people are in close interaction. To address this, we propose a novel method leveraging the personalized implicit neural avatar of each individual as a prior, which significantly improves the robustness and precision of this challenging pose estimation task. Concretely, the avatars are efficiently reconstructed via layered volume rendering from sparse multi-view videos. The reconstructed avatar prior allows for direct optimization of 3D poses based on color and silhouette rendering loss, bypassing the issues associated with noisy 2D detections. To handle interpenetration, we propose a collision loss on the overlapping shape region of avatars to add penetration constraints. Moreover, both 3D poses and avatars are optimized in an alternating manner. Our experimental results demonstrate state-of-the-art performance on several public datasets.

Keywords: human pose estimation · human close interaction · multi-view pose estimation · avatar prior

1 Introduction

People frequently engage in interactions with each other in daily life to offer physical support or convey emotions. For AI systems to interpret 3D human interactions, a foundational step is to reconstruct 3D human poses and shapes using consumer-grade sensors like cameras. However, in closely interacting scenarios, such as hugging or kissing, 3D pose estimations face challenges due to strong occlusions. To predict 3D poses from 2D images or videos, the problem of depth ambiguity gets worse when close contact is involved. Consequently, a multi-camera setup becomes essential to offer additional observations and to address depth ambiguity in pose estimation.

Despite the ubiquity of close human interactions, the study of estimating such human motions is scarce. Most previous multi-human methods with multi-view setup [2–4,19,22,30,57,65,71] focus on scenarios where people are at a distance from each other. Some methods [2, 4, 19, 22, 71] typically formulate the problem as a cross-view matching problem, relying heavily on 2D joint estimations for subsequent 3D triangulation. These methods demonstrate high sensitivity to noisy or missing 2D joint estimations, particularly when occlusion occurs. Another group of learning-based methods [57, 60, 62, 65] attempts to integrate 2D features from each view into a 3D voxel space and predict 3D human poses from identified 3D subvolumes of each individual. These methods are more robust to occlusion, but they encounter challenges with generalization and are sensitive to distribution shifts due to the lack of annotated 3D data. To correct abnormal or missing pose estimations, some methods [18, 21] leverage parametric body models like SMPL [37] as full-body priors [18, 21] and fit these 3D models to 2D joint estimations. Although this method alleviates the issue of abnormal or missing joint estimations, it remains constrained by noisy 2D detections.

Embracing the challenging problem of pose estimation for close interactions, we propose a novel method to estimate the 3D poses and shapes of multiple people, observed from a sparse set of cameras. Our goal is to ensure that even in close contact, the estimated human poses and shapes are accurate and without interpenetrations. The key of our method is reconstructing implicit textured avatars of each individual in the scene and leveraging them as a strong personalized prior for pose optimization (Figure 1). In contrast to relying solely on noisy 2D joint detections, this textured avatar prior enables us to leverage color and silhouette information for pose refinement, which significantly increases the robustness and accuracy of our method (Table 3). Meanwhile, the reconstructed avatar provides crucial geometric and appearance information to avoid collisions between individuals. Compared to methods using SMPL body shape [25, 42] to penalize collisions, our implicit avatar model contains a more detailed clothed shape and additional appearance cues and enables efficient computation of penetration loss. Due to the mutually beneficial relationship between avatar and pose, we alternate between pose optimization and avatar refinement.

More specifically, to accelerate the learning of avatars, we model each human individual in canonical space using an efficient neural radiance field variant in Instant NGP [43] and combine it with an efficient SMPL-based deformation

module [37]. To learn and render avatar models of multiple people, we adapt layered volume rendering [52, 69] to our avatar model. This adaptation allows us to jointly optimize all avatar models through a straightforward rendering loss. Once learned, the avatar can be animated and rendered based on pose parameters at interactive rates, thus naturally serving as an efficient personalized textured prior for pose optimization. With the learned personalized prior, we optimize pose via a novel objective function. Different from previous methods based on 2D reprojection error of joints, we directly optimize pose parameters via minimizing color and silhouette rendering losses while keeping the learned avatar model fixed. To prevent interpenetration between human individuals, a collision loss is introduced by penalizing the situation when a 3D point is occupied by multiple avatars. To remove artifacts in the initial avatar due to imperfect pose initialization, we further refine avatars based on optimized poses. Throughout the optimization process, both initial personalized avatar models and SMPL parameters are optimized in an alternating manner, motivated by the insight that accurate 3D pose estimations improve avatar learning, and improved avatar models, in turn, increase the precision of overall pose estimations.

We experimentally demonstrate that our method significantly outperforms previous state-of-the-art methods on several public datasets both quantitatively (Table 1 and Table 2) and qualitatively (Figure 3) especially when people are in close interaction. In summary, our contributions are:

- We propose a pipeline that efficiently creates implicit neural avatars of closely interacting people and leverages the learned avatars as priors to optimize poses.
- The avatar prior enables us to design a novel objective function that leverages color and mask rendering loss for pose optimization. We show the superiority of this loss function compared to the 2D reprojection error of 3D joints, which is used by most of the previous methods.
- Based on the learned avatar, a collision loss is introduced to avoid penetration when individuals are in contact.

2 Related Work

2.1 Multi-Person 3D Pose Estimation

Despite significant progress in multi-human 2D pose estimation [9, 13, 15, 24, 34, 48] and 3D pose estimation from monocular image or video [7, 31–33, 38–40, 45, 55, 56, 58, 67, 68], the reconstruction accuracy is still limited due to depth ambiguity and strong occlusions when humans are in close contact with each other. A multi-view setting [4, 5, 19, 22, 23, 35, 36, 53, 60, 62, 65, 71] helps to alleviate these challenges. One straightforward idea of most methods [2, 4, 19, 22, 71, 76] is to formulate the problem into cross-view matching and association problems. MVPose [19] performs 2D person parsing in each image and leverages cross-view person matching to infer 3D pose. 4DAssociation [71] additionally adds tracking

into this process to form a unified graph for associating 4D information. However, these methods are sensitive to the noisy estimation of 2D pose. In contrast to these matching-based methods, some recent methods [5, 14, 17, 23, 35, 49, 57, 60, 62, 65, 70, 75] directly learn deep neural networks to regress poses. Faster Voxelpose [65] employs the feature volume proposed by Voxelpose [57] and enhances computational efficiency. Graph [62] designs three graph neural network models for human center detection and pose estimation. MvP [60] simplifies the multi-person pose estimation by direct regression using the transformer model. A concurrent method CloseMocap [53] proposes to learn a model from a synthetic dataset simulating occlusion situations. However, heavily relying on the 2D or 3D features as input during training, these learning-based methods suffer from generalization issues when subjects, motions, and camera configurations change.

To further improve the robustness, a statistical parametric body model such as SMPL [37] is explored in [21] as a regularization prior for 3D joint refinement. Some follow-ups [18, 72] show that parametric models help in correcting implausible 3D pose estimates and filling in missing joints. However, this coarse body prior highly relies on aligning the 3D joints to 2D pose estimations, which are inaccurate when occlusion happens. Different from all of these previous methods, we explore the usage of personalized textured avatar models as priors to refine human poses. This prior enables us to leverage color and silhouette information from multi-view observation for refining poses of closely interacting humans.

2.2 3D Human Modeling

Parametric human body models [1, 28, 37, 44, 64] can represent minimally clothed human shapes by deforming a template mesh. It is challenging to extend this explicit representation for modeling clothed humans due to the fixed topology and resolution. To overcome this limitation, methods such as SNARF [12] and SCANimate [51] propose to model articulated human shapes based on 3D implicit representations. Many works [6, 8, 10, 20, 26, 29, 47, 59, 61] fit implicit neural fields to RGB or RGB-D videos by neural rendering to reconstruct the shape and appearance of a single human body. However, when applied to a multi-human scene, these methods are not able to achieve good fidelity due to strong occlusions. Recent methods including ST-NeRF [69] and [52] leverage layered neural representation to model multiple humans with sparse multi-view videos and thus can generate novel view synthesis of dynamic multiple humans. The main problem of all aforementioned approaches, however, is their high reliance on 3D human pose estimation [69, 73]: the deformation of the human model requires accurate human poses, which is hard to obtain when people are in close interaction. To address this challenge, our method is orthogonal to others, aiming at leveraging the learned avatar as priors for pose estimation in close interaction.

2.3 3D Human Datasets for Close Interaction

Most of the existing datasets [3, 11, 30] like Shelf and Panoptic studying multi-person pose estimation focus on the scene where people are at a distance from

each other and rarely involved in close interactions. To study close interactions among people, MultiHuman [73] dataset captures multi-person interaction with some close interactions and occlusions. ExPI [27] creates a multi-person extreme motion dataset with close interactions, but it focuses mainly on motion prediction for future frames instead of pose estimation from sparse views. CHI3D [25] captures two-person interaction datasets and proposes to learn a contact estimation module from annotations to improve the precision of pose estimation. The most recent work Hi4D [66] creates a challenging dataset of physically close human interaction and proposes a method to disentangle human bodies and estimate poses. However, this method relies on the 3D ground truth of clothed human meshes captured with expensive 3D body scanners. Due to the reliance on the limited annotated 3D data, this method faces challenges in generalization with different people and camera configurations. In contrast to all the works before, we intend to solve the problem of pose estimation in close human interaction without requiring accurate 3D scans or other training data.

3 Method

Given a dynamic scene captured by a sparse set of RGB cameras, our goal is to estimate the 3D pose and shape of multiple people even if they interact closely. To address this challenging task, our key idea is to first reconstruct the personalized avatar of each individual in the scene and leverage them as a strong prior to refine the appearance and pose in an alternating manner. An overview of our method is shown in Figure 2.

We first introduce an efficient pipeline to create avatars of multiple people in a scene (Section 3.1 and Figure 2(a)). Specifically, we leverage an accelerated neural radiance field to represent the shape and appearance of each individual in canonical space and deform it at an interactive rate. We then adapt layered volume rendering to our pipeline, which composites the rendering of avatars into one image, thus enabling direct learning from multi-view video inputs.

Thanks to the learned avatar prior of each individual, we can enhance 3D pose optimization via a combination of RGB and silhouette rendering loss (Section 3.2 and Figure 2(b)). While previous work highly relies on noisy 2D joint detection, we show that employing such color and silhouette information can largely increase precision and robustness. Moreover, a collision loss is introduced to avoid interpenetration. Finally, we alternate between avatar learning and pose optimization to generate complete and accurate 3D human poses.

3.1 Multi-Avatar Prior Learning

Avatar Model We represent each human individual in canonical space using an accelerated neural radiance field [43] and model shape-aware articulated deformation based on SMPL [37].

– **Canonical Appearance Representation:** To model human shape and appearance, we create canonical radiance field $\bar{\mathbf{F}}_{\sigma_f}^{(l)}$ for each human instance

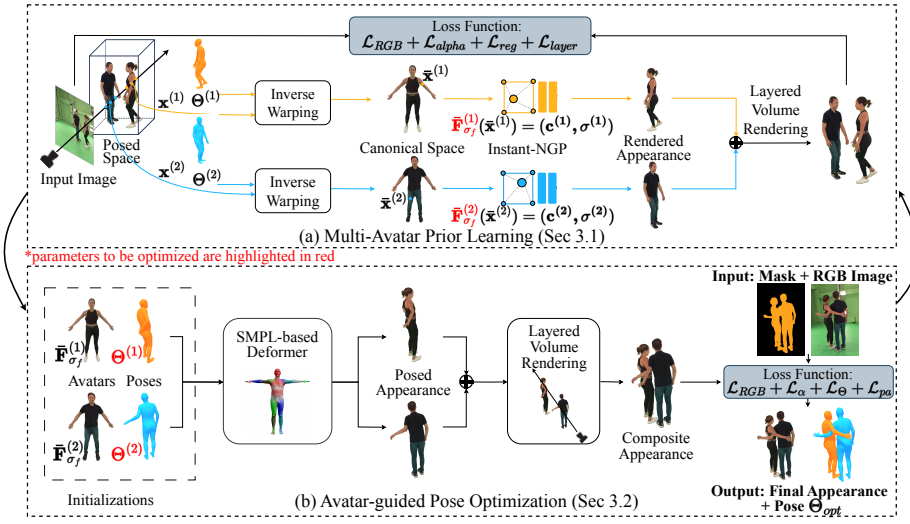

Fig. 2: Method Overview: Our method consists of two modules: *(a) Multi-Avatar Prior Learning:* Given the input multi-view images and estimated poses $\mathbf{\Theta}^{(l)}$, we sample points $\mathbf{x}^{(l)}$ for each individual l along the rays in posed space and warp these points into canonical space and calculate their color $\mathbf{c}^{(l)}$ and density $\sigma^{(l)}$ via the canonical appearance network $\bar{\mathbf{F}}^{(l)}_{\sigma_f}$. We leverage layered volume rendering [69] to attain the final pixel color and compare it with the original input image to optimize the parameters of avatars. *(b) Avatar-guided Pose Optimization:* Given learned avatar model $\bar{\mathbf{F}}^{(l)}_{\sigma_f}$ and initial poses $\mathbf{\Theta}^{(l)}$ of each individual l, we deform all of the avatars based on SMPL-based deformer and render them jointly via layered volume rendering. We compare the composite rendering with input observation and minimize the RGB and mask rendering loss to optimize poses. A collision loss is additionally introduced to avoid interpenetration. Finally, we alternate between two modules to obtain the final result. For clarity, the parameters to be optimized are marked as red in each module.

$l \in [1, L]$, where L is the number of humans in the scene. $\bar{\mathbf{F}}^{(l)}_{\sigma_f}$ takes a 3D point $\bar{\mathbf{x}}^{(l)}$ as input and predicts its density $\sigma^{(l)}$ and color $\mathbf{c}^{(l)}$. Following Instant-NGP [43] and InstantAvatar [29] to accelerate the rendering, $\bar{\mathbf{F}}^{(l)}_{\sigma_f}$ is parameterized via using a hash table to store feature grids at different scales.

- **Pose Representation**: We represent the 3D pose and underlying body shape for all human instances by SMPL parameters $\mathbf{\Theta} = \{\mathbf{\Theta}^{(l)}\}_{l\in[1,L]}$. For each human l, $\mathbf{\Theta}^{(l)} = \{\beta^{(l)}, \theta^{(l)}, t^{(l)}\}$ contains shape parameters $\beta^{(l)} \in \mathbb{R}^{10}$, pose parameters $\theta^{(l)} \in \mathbb{R}^{72}$ and translation $t^{(l)} \in \mathbb{R}^3$ of SMPL.
- **Deformer:** To enable animation given targeted poses $\mathbf{\Theta}^{(l)}$, we require the radiance field in the posed space. Given a point $\mathbf{x}^{(l)}$ in deformed space of human l, we determine the corresponding canonical point $\bar{\mathbf{x}}^{(l)}$ by inverse linear blend skinning(LBS) [37]: $\bar{\mathbf{x}}^{(l)}(\mathbf{x}^{(l)}, \mathbf{\Theta}^{(l)}) = (\sum_{i=1}^{n_b} w_i(\mathbf{\Theta}^{(l)})\mathbf{B}_i(\mathbf{\Theta}^{(l)}))^{-1}\mathbf{x}^{(l)}$, where $\mathbf{B}_i$ is the rigid bone transformation matrix for joint $i \in \{1, ..., n_b\}$ under pose $\mathbf{\Theta}^{(l)}$. w_i is the skinning weights of the nearest neighbor of $\mathbf{x}^{(l)}$ in the deformed SMPL vertices. We obtain the radiance field at the point $\mathbf{x}^{(l)}$

by evaluating the canonical appearance field at the corresponding point $\bar{\mathbf{x}}^{(l)}$.

Layered Volume Rendering To obtain the pixel value for a ray $\mathbf{r} \in \mathcal{R}$, we raycast every human instance separately with a layered rendering strategy similar to ST-NeRF [69]. Specifically, we first calculate the intersection points between the ray and the 3D bounding box of each human instance and uniformly sample points in each bounding box. To distinguish different identities, we assign each sampled point $\mathbf{x}_i$ a one-hot representation $\mathbf{m}_i = [m_i^{(1)}, \cdots, m_i^{(L)}]$ to indicate which human identity it belongs to. After sorting all sampled points by their depth values and calculating their corresponding color $\mathbf{c}_i$ and density σ_i from the avatar model, if $m_i^{(l)} = 1$, we compute

$$\mathbf{c}_i, \sigma_i = \bar{\mathbf{F}}_{\sigma_f}^{(l)}(\bar{\mathbf{x}}_i(\mathbf{x}_i, \mathbf{\Theta}^{(l)})). \tag{1}$$

The color of each ray is computed via numerical integration [41].

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^{N} \alpha_i \prod_{j<i} (1 - \alpha_j)\mathbf{c}_i \quad \alpha_i = 1 - \exp(-\sigma_i \delta_i), \tag{2}$$

where δ_i is the distance between samples. The accumulated alpha value, which represents ray opacity, can be computed via:

$$\alpha(\mathbf{r}) = \sum_{i=1}^{N} \alpha_i \prod_{j<i} (1 - \alpha_j). \tag{3}$$

For each human identity $l \in [1, L]$, the corresponding instance ray opacity can be calculated via:

$$\alpha^{(l)}(\mathbf{r}) = \sum_{i=1}^{N} \alpha_i \prod_{j<i} (1 - \alpha_j) m_i^{(l)}. \tag{4}$$

Training The overall training process is shown in Figure 2(a). For training avatar layers, we minimize the Huber loss ρ between the predicted pixel color $\hat{\mathbf{C}}(\mathbf{r})$ and the ground truth pixel color $\mathbf{C}_{gt}(\mathbf{r})$:

$$\mathcal{L}_{RGB} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \rho(\|\hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}_{gt}(\mathbf{r})\|). \tag{5}$$

Since instance segmentation of human performers is hard to obtain and is not accurate, we choose foreground segmentation as our mask supervision, which is obtained via SAM-Track [16]. We apply a loss for optimizing the rendered alpha values α to reduce the artifacts in the floating area:

$$\mathcal{L}_{alpha} = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} (\alpha(\mathbf{r}) - \alpha_{SAM}(\mathbf{r}))^2. \tag{6}$$

Following [52], we add a regularization loss for instance alpha values to make sure every pixel can only be rendered from one human layer:

$$\mathcal{L}_{layer} = -\frac{1}{L \mid \mathcal{R} \mid} \sum_{\mathbf{r} \in \mathcal{R}} \sum_{l=1}^{L} \alpha^{(l)}(\mathbf{r}) \log(\alpha^{(l)}(\mathbf{r})). \tag{7}$$

Similar to [29], we also add hard surface and density regularization terms in the learning process. More training details can be found in the Supp Mat.

3.2 Avatar-guided Pose Optimization

Equipped with learned avatars obtained in Section 3.1, we aim to estimate the 3D shape and pose of multiple humans with close physical contact. To achieve this, we leverage the avatars as priors to handle the challenging pose ambiguities caused by contact. Specifically, we first initialize pose parameters using an off-the-shelf 3D pose estimator [71] and refine the pose via a rendering loss between rendered posed avatars and 2D observations (Figure 2(b)). Since the initial imperfect pose estimations may cause artifacts in avatar reconstruction, we further refine the weights of the avatar model using the optimized pose. This process is formulated as an alternating optimization to refine both poses and avatars.

Initialization The initial 3D pose proposals are estimated by leveraging the off-the-shelf 3D human pose estimator [71]. After that, we register a SMPL model to the estimated 3D joints to obtain the initial pose parameters $\mathbf{\Theta_0}$. Given these estimated poses, the initial avatar model of each individual is further learned from multi-view videos.

Objective To leverage avatars as priors to tackle challenges caused by contact, we optimize the following objective:

$$\begin{aligned} \mathcal{L}(\mathbf{\Theta}) = \lambda_{RGB}\mathcal{L}_{RGB}(\mathbf{\Theta}) + \lambda_{\alpha}\mathcal{L}_{\alpha}(\mathbf{\Theta}) \\ + \lambda_{reg}\mathcal{L}_{reg}(\mathbf{\Theta}) + \lambda_{pa}\mathcal{L}_{pa}(\mathbf{\Theta}). \end{aligned} \tag{8}$$

Here, $\mathbf{\Theta}$ is the SMPL parameters to be optimized for all avatars, which is also consistent with the pose parameters to deform the avatar model (Section 3.1). Following [61], we represent $\mathbf{\Theta}$ as $\mathbf{\Theta_0} + \text{MLP}(\mathbf{\Theta_0})$ and refine poses by changing the parameters of the neural network. This representation empirically shows more robust results compared to directly optimizing SMPL parameters.

Different from previous methods [18,21,57,71], the personalized multi-avatar prior allows us to leverage appearance and silhouette information to refine initial poses. Specifically, we calculate $\mathcal{L}_{RGB}$ to ensure the color consistency between the rendered pixel $\hat{\mathbf{C}}(\mathbf{r}, \mathbf{\Theta})$ (Equation (1) and Equation (2)) of deformed avatar with poses $\mathbf{\Theta}$ and the corresponding ground-truth pixel color $\mathbf{C}_{gt}(\mathbf{r})$.

$$\mathcal{L}_{RGB}(\mathbf{\Theta}) = \frac{1}{\mid \mathcal{R} \mid} \sum_{r \in \mathcal{R}} \rho(\|\hat{\mathbf{C}}(\mathbf{r}, \mathbf{\Theta}) - \mathbf{C}_{gt}(\mathbf{r})\|). \tag{9}$$

Additionally, a cross-entropy loss $\mathcal{L}_\alpha$ is introduced to ensure that the rendered mask $\alpha(\mathbf{r}, \mathbf{\Theta})$ (Equation (3)) of the reposed avatar is aligned with the estimated SAM-Track mask $\alpha_{SAM}(\mathbf{r})$ by:

$$\mathcal{L}_\alpha(\mathbf{\Theta}) = -\sum_{\mathbf{r}\in\mathcal{R}} \alpha_{SAM}(\mathbf{r}) \log(\alpha(\mathbf{r}, \mathbf{\Theta})) \tag{10}$$

To penalize the unnatural poses and avoid elbows and knees bending in the wrong direction, we add an L2 regularization term and combine it with the pose prior in SMPLify [7] to constrain physically implausible joint rotation:

$$\mathcal{L}_{reg}(\mathbf{\Theta}) = \|\mathbf{\Theta}\|_2 + \lambda \sum_{i\in I} \exp(\mathbf{\Theta}_i), \tag{11}$$

where I is the set of pose indices corresponding to elbows and knees.

A key challenge to correctly estimate poses in close interaction is to handle interpenetration. Since every avatar is modeled separately, the surfaces tend to intersect when they are in contact. To handle this, we first select sampled points inside multiple instances as $\mathcal{S} = \{\mathbf{x}_i \mid \alpha_i^{(p)} > 0, \alpha_i^{(q)} > 0, p, q \in [1, L], p \neq q\}$ ($\alpha_i^{(p)}$ is calculated from Equation (1) and Equation (2) corresponding with a point $\mathbf{x}_i$ with $m_i^{(p)} = 1$). We then propose a collision loss $\mathcal{L}_{pa}$ for penalizing penetration:

$$\mathcal{L}_{pa}(\mathbf{\Theta}) = \frac{1}{\mid \mathcal{S} \mid} \sum_{\mathbf{x}_i\in\mathcal{S}} \alpha_i^{(p)}(\mathbf{\Theta})\alpha_i^{(q)}(\mathbf{\Theta}) \tag{12}$$

Intuitively, this loss guarantees every sample point in 3D space can not be occupied by multiple avatar models simultaneously, which guarantees better pose estimation in close contact.

Alternating Optimization Since artifacts sometimes appear on initial avatars due to imperfect pose initialization, we further refine avatars based on optimized poses via minimizing the loss function in Section 3.1. Finally, the optimization of poses and avatars is formulated in an alternating fashion for N steps. More details of optimization can be found in the Supp Mat.

4 Experiments

Datasets. We mainly evaluate our proposed method on Hi4D [66] and CHI3D [25] Dataset, which are challenging datasets of two humans in close interaction. To further demonstrate the generalization ability of our method for more than two people, we also evaluate our method on Shelf [3] and MultiHuman [52] Dataset, which includes three or four people. More details are shown in the Supp Mat.

Metrics. We use the Mean Per Joint Position Error (MPJPE) [57] to measure the distance between the ground truth 3D poses and the estimated poses. We also choose the Percentage of Correct Parts (PCP3D) metric [19] to calculate the percentage of correct parts. AP_K [57] and Recall [21] are also leveraged to evaluate performance. More details are shown in the Supp Mat.

Baselines. We compare our method on the Hi4D dataset with state-of-the-art methods in three categories discussed in Section 2.1. For learning-based methods, we choose Graph [63], MvP [60], and Faster VoxelPose [65] and fine-tune these models on a subset of the evaluating datasets. For pure association-based methods, we choose 4DAssociation [71] and MVPose* [19]. Based on initial results, MVPose [18] added temporal tracking and SMPL prior to MVPose* [19], which is regarded as an SMPL-guided method. More details about baselines can be found in the Supp Mat.

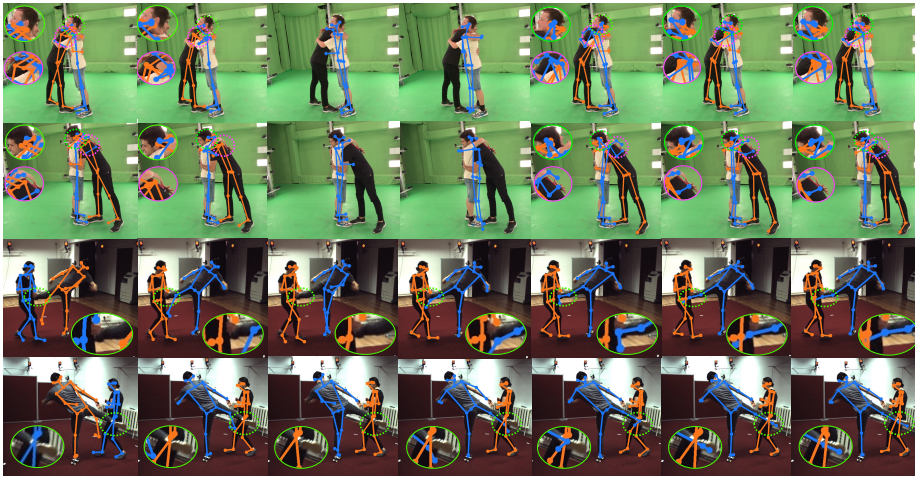

Fig. 3: Qualitative Comparison with SotA methods [19,60,63,65,71] on Hi4D and CHI3D. We show two examples from the Hi4D and CHI3D datasets compared with Graph, MvP, Faster VoxelPose, MVPose, and 4DAssociation. For each example, we show 2D projections on two sampled views.

4.1 Comparison to SotA

Table 1 and Table 2 summarizes our quantitative comparisons on Hi4D and CHI3D with SotA (State-of-the-art) methods. Our method largely outperforms the other SotA methods in all of the metrics including MPJPE, PCP3D, and AP_K. More comparisons on Shelf and MultiHuman are shown in the Supp Mat.

Comparison with Graph [62], MvP [60], Faster VoxelPose [65]. Our method achieves much better quantitative results compared to Graph, MvP, and Faster VoxelPose on Hi4D and CHI3D. These learning-based methods are prone to overfitting to training pose distribution, resulting in an inability to accurately estimate the challenging poses of interacting actors, such as touching shoulders or legs. More specifically, Faster VoxelPose sometimes causes missing actors in close contact, leading to a relatively low recall. Graph and MvP perform better

Method	MPJPE(mm) ↓	PCP(%) ↑	AP_{50} ↑	AP_{100} ↑	Recall(%) ↑
MvP [60]	92.77	74.14	41.59	63.86	93.84
Graph [62]	89.62	71.55	44.75	67.33	93.31
Faster VoxelPose [65]	68.40	73.67	44.05	68.70	83.55
MVPose* [19]	53.05	87.57	67.97	80.28	93.80
MVPose [18]	42.63	90.76	71.79	90.19	93.30
4DAssociation [71]	41.29	88.62	80.87	97.27	**98.78**
Ours	**32.10**	**96.90**	**91.48**	**97.33**	**98.78**

Table 1: Quantitative Comparison with SotA on the Hi4D [66] Dataset (8 views). We compare our method with MvP [60], Graph [62], Faster VoxelPose [65], MVPose* [19], MVPose [18] and 4DAssociation [71]. We report MPJPE, PCP, AP_K, and Recall metric for all methods.

Method	MPJPE(mm) ↓	PCP(%) ↑	AP_{50} ↑	AP_{100} ↑	Recall(%) ↑
MvP [60]	55.38	89.47	63.58	92.53	99.06
Graph [62]	45.33	92.46	74.02	95.25	99.17
Faster VoxelPose [65]	67.81	78.41	29.28	82.88	93.34
MVPose* [19]	50.42	90.39	69.13	75.72	88.72
MVPose [18]	34.05	93.35	79.94	86.91	88.18
4DAssociation [71]	37.47	99.30	89.66	98.67	**99.85**
Ours	**32.98**	**99.79**	**93.20**	**99.79**	**99.85**

Table 2: Quantitative Comparison with SotA on the CHI3D [25] Dataset (4 views). We compare our method with Faster VoxelPose [65], MVPose* [19], MVPose [18] and 4DAssociation [71]. We report MPJPE, PCP, AP_K, and Recall metric for all methods.

in recall, but Graph fails to consistently track the actors across frames, and MvP results in many misaligned joints between actors.

Comparison with MVPose [18]. Compared with MVPose, the MPJPE and precision of our method are much better. Specifically, when the threshold of precision becomes smaller, the gap of precision becomes larger. This is because the precision of fitting the SMPL body heavily relies on 2D joint detections, which are noisy and inaccurate when close interaction happens. In contrast, our method takes full advantage of color and silhouette rendering loss to optimize poses, leading to robustness to occlusions. Furthermore, this top-down method also cannot detect closely interacting actors correctly due to strong occlusions shown in Figure 3.

Comparison with 4DAssociation [71]. Finally, we compare our method with the bottom-up association method. As shown in Table 1 and Table 2, our method outperforms 4DAssociation in most metrics. Figure 3 shows that when actors are close, this bottom-up method is inclined to associate joints with the wrong human instances. This is because they solve the joint association with a greedy algorithm, which is sensitive to missing and inaccurate 2D joint detections.

4.2 Additional Qualitative Samples

Figure 4 demonstrates more qualitative results of our method on Hi4D, CHI3D, and MultiHuman Real-Cap with challenging and close interactions among 2 or 3 people. we also demonstrate results on the Shelf [3] dataset, which contains 4 people without close contact. More results can be found in the Supp Mat.

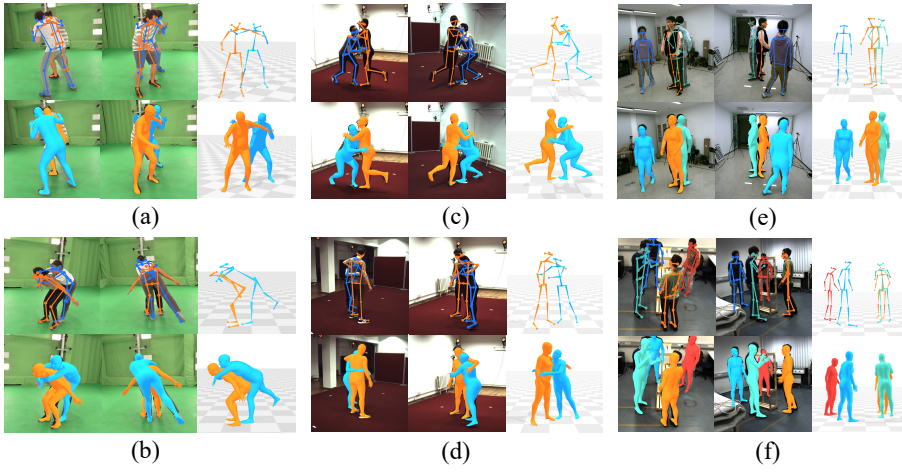

Fig. 4: Qualitative Results of our method on Hi4D (a)(b), CHI3D (c)(d), MultiHuman Real-Cap (e), and Shelf (f). The left and middle columns in each sub-figure show the 2D projections of the estimated 3D skeletons and SMPL body meshes on two views. The right column in each sub-figure demonstrates skeletons and SMPL bodies in 3D scenes.

4.3 Ablation Study

To validate the effectiveness of our method, we conduct a detailed analysis of different design choices of our algorithm. All the experiments are conducted on the Hi4D Dataset.

Comparison with SMPL Body Prior. To validate the effectiveness of our personalized avatar prior, we compare our method with a baseline that optimizes SMPL parameters to align reprojections of 3D joints to 2D observations. As shown in Table 3, our method outperforms the SMPL prior baseline by a significant margin. This is due to fitting errors when 2D pose estimation is not accurate. In contrast, thanks to the learned avatar prior, our method can leverage color and silhouette rendering loss to refine poses, without high reliance on joint detection and alignment which are not accurate when occlusion happens. Figure 5 shows a qualitative comparison. We observe that the pose of the arm in the baseline method is wrongly estimated and even leads to penetration, whereas our method still reconstructs the poses correctly.

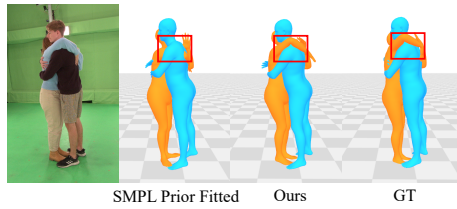

Fig. 5: Comparison with SMPL Body Prior. Only fitting SMPL to 2D observations, some joints in close contact such as arms are incorrectly estimated and even cause intersections between body surfaces. In contrast, our personalized prior enables accurate estimation of poses.

Color and Mask Loss. To demonstrate the effectiveness of the color and silhouette rendering loss in our optimization process, we design baselines without RGB loss and mask loss respectively for comparison. Table 3 shows that the RGB loss significantly improves pose optimization, and the optimization process will completely deviate from the correct trajectory in the absence of RGB loss. Mask loss is proven to slightly increase the accuracy by adding additional constraints on the rendered human silhouette.

Method	**MPJPE**(mm) ↓	**PCP**(%) ↑	**AP**$_{50}$ ↑	**AP**$_{100}$ ↑
Ours (SMPL fitted)	40.41	95.10	84.04	94.37
Ours (w/o RGB loss)	78.40	83.26	17.84	74.55
Ours (w/o Mask loss)	31.00	97.33	90.28	99.06
Ours (Joint Optimization)	66.04	84.56	22.55	76.70
Ours	**29.37**	**98.02**	**96.79**	**99.06**

Table 3: Quantitative Ablation Results. Ablations to evaluate our method with only the SMPL fitted method, our method without RGB loss and without Silhouette loss, and our method without alternating optimization.

Alternating Optimization. To show the advantage of alternating optimization, a joint optimization is selected for comparison, which is widely used in avatar reconstruction [20, 26]. In Table 3, we observe that our method largely outperforms the baseline. As shown by the qualitative results in Figure 6, the final avatar of our ablated baselines suffers from artifacts and floating points around contact body parts. This imperfect avatar in turn causes the wrong 3D pose estimations. In contrast, our method can faithfully reconstruct avatars and poses under challenging poses.

Penetration Loss. Our penetration loss serves an important role in avoiding interpenetration. Comparing our method to the ablated version where we remove this loss, Figure 7 shows that the SMPL body of one person partially intersects with the other person in the contact area. This is also confirmed by the rendering result of the avatar. By penalizing the collision of density fields of avatars in

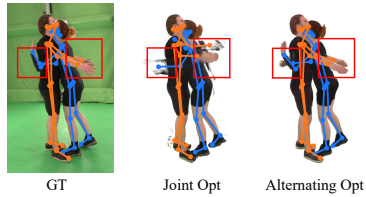

Fig. 6: Ablation of Alternating Optimization. We show the results of rendered avatars and projections of the estimated 3D poses. Joint optimization suffers from artifacts around the contact part and in turn causes wrong pose estimations. In contrast, ours reconstructs both avatars and poses correctly.

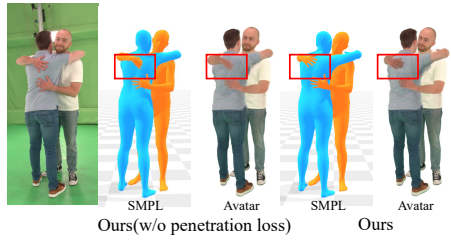

Fig. 7: Ablation of Penetration Loss. Without penetration loss, both the avatar and underlying SMPL body tend to have collisions on surfaces.

3D space, our method largely reduces the penetration and thus achieves more accurate pose estimation.

4.4 Limitations and Future Work

We do not model hands in our avatar model, it will be a promising direction to integrate hand models [46, 50] into the personalized avatar. The current optimization pipeline is not fast enough, which can be accelerated with more efficient representation [74] or more powerful optimization tools [54]. We believe it will be interesting future work to adapt our pipeline to estimate 3D poses of closely interacting people from in-the-wild and monocular videos and images. More discussions about limitations and future work can be found in the Supp Mat.

5 Conclusion

In this paper, we propose AvatarPose, a novel method to estimate the 3D poses of multiple people in close interaction from sparse multi-view videos. Unlike previous methods leveraging SMPL body prior, we first reconstruct the avatar of each individual and leverage the avatars as personalized priors to guide pose optimization. The avatar prior enables us to use color and silhouette observations, instead of relying on noisy 2D joint detections. A collision loss is also introduced to constrain penetration in close contact. Our method outperforms SotA methods significantly on public datasets of close human interactions. The code associated with this paper will be released upon acceptance.

References

1. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM SIGGRAPH 2005 Papers, pp. 408–416 (2005) 4
2. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures for multiple human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1669–1676 (2014) 2, 3
3. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures for multiple human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1669–1676 (2014) 2, 4, 9, 12
4. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures revisited: Multiple human pose estimation. IEEE transactions on pattern analysis and machine intelligence **38**(10), 1929–1942 (2015) 2, 3
5. Benzine, A., Chabot, F., Luvison, B., Pham, Q.C., Achard, C.: Pandanet: Anchor-based single-shot multi-person 3d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6856–6865 (2020) 3, 4
6. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Combining implicit function learning and parametric models for 3d human reconstruction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 311–329. Springer (2020) 4
7. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 561–578. Springer (2016) 3, 9
8. Burov, A., Nießner, M., Thies, J.: Dynamic surface function networks for clothed human bodies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10754–10764 (2021) 4
9. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299 (2017) 3
10. Chen, J., Zhang, Y., Kang, D., Zhe, X., Bao, L., Jia, X., Lu, H.: Animatable neural radiance fields from monocular rgb videos. arXiv preprint arXiv:2106.13629 (2021) 4
11. Chen, L., Ai, H., Chen, R., Zhuang, Z., Liu, S.: Cross-view tracking for multi-human 3d pose estimation at over 100 fps. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3279–3288 (2020) 4
12. Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11594–11604 (2021) 4
13. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7103–7112 (2018) 3
14. Chen, Y., Gu, R., Huang, O., Jia, G.: Vtp: volumetric transformer for multi-view multi-person 3d pose estimation. Applied Intelligence pp. 1–12 (2023) 4

15. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5386–5395 (2020) 3
16. Cheng, Y., Li, L., Xu, Y., Li, X., Yang, Z., Wang, W., Yang, Y.: Segment and track anything. arXiv preprint arXiv:2305.06558 (2023) 7
17. Choudhury, R., Kitani, K.M., Jeni, L.A.: Tempo: Efficient multi-view pose estimation, tracking, and forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14750–14760 (2023) 4
18. Dong, J., Fang, Q., Jiang, W., Yang, Y., Bao, H., Zhou, X.: Fast and robust multi-person 3d pose estimation and tracking from multiple views. In: T-PAMI (2021) 2, 4, 8, 10, 11
19. Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X.: Fast and robust multi-person 3d pose estimation from multiple views. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7792–7801 (2019) 2, 3, 9, 10, 11
20. Dong, Z., Guo, C., Song, J., Chen, X., Geiger, A., Hilliges, O.: Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20470–20480 (2022) 4, 13
21. Dong, Z., Song, J., Chen, X., Guo, C., Hilliges, O.: Shape-aware multi-person pose estimation from multi-view images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11158–11168 (2021) 2, 4, 8, 9
22. Ershadi-Nasab, S., Noury, E., Kasaei, S., Sanaei, E.: Multiple human 3d pose estimation from multiview images. Multimedia Tools and Applications **77**, 15573–15601 (2018) 2, 3
23. Fabbri, M., Lanzi, F., Calderara, S., Alletto, S., Cucchiara, R.: Compressed volumetric heatmaps for multi-person 3d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7204–7213 (2020) 3, 4
24. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 2334–2343 (2017) 3
25. Fieraru, M., Zanfir, M., Oneata, E., Popa, A.I., Olaru, V., Sminchisescu, C.: Three-dimensional reconstruction of human interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7214–7223 (2020) 2, 5, 9, 11
26. Guo, C., Jiang, T., Chen, X., Song, J., Hilliges, O.: Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12858–12868 (2023) 4, 13
27. Guo, W., Bie, X., Alameda-Pineda, X., Moreno-Noguer, F.: Multi-person extreme motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13053–13064 (2022) 5
28. Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., Seidel, H.P.: A statistical model of human pose and body shape. In: Computer graphics forum. vol. 28, pp. 337–346. Wiley Online Library (2009) 4
29. Jiang, T., Chen, X., Song, J., Hilliges, O.: Instantavatar: Learning avatars from monocular video in 60 seconds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16922–16932 (2023) 4, 6, 8

30. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3334–3342 (2015) 2, 4
31. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7122–7131 (2018) 3
32. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: Part attention regressor for 3d human body estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11127–11137 (2021) 3
33. Kocabas, M., Huang, C.H.P., Tesch, J., Müller, L., Hilliges, O., Black, M.J.: Spec: Seeing people in the wild with an estimated camera. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11035–11045 (2021) 3
34. Li, Y., Zhang, S., Wang, Z., Yang, S., Yang, W., Xia, S.T., Zhou, E.: Tokenpose: Learning keypoint tokens for human pose estimation. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 11313–11322 (2021) 3
35. Lin, J., Lee, G.H.: Multi-view multi-person 3d pose estimation with plane sweep stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11886–11895 (2021) 3, 4
36. Liu, Y., Gall, J., Stoll, C., Dai, Q., Seidel, H.P., Theobalt, C.: Markerless motion capture of multiple characters using multiview image segmentation. IEEE transactions on pattern analysis and machine intelligence **35**(11), 2720–2735 (2013) 3
37. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015) 2, 3, 4, 5, 6
38. Luvizon, D.C., Habermann, M., Golyanik, V., Kortylewski, A., Theobalt, C.: Scene-aware 3d multi-human motion capture from a single camera. In: Computer Graphics Forum. vol. 42, pp. 371–383. Wiley Online Library (2023) 3
39. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 2640–2649 (2017) 3
40. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. Acm transactions on graphics (tog) **36**(4), 1–14 (2017) 3
41. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 7
42. Muller, L., Osman, A.A.A., Tang, S., Huang, C.H.P., Black, M.J.: On self-contact and human pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9990–9999 (June 2021) 2
43. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022) 2, 5, 6
44. Osman, A.A., Bolkart, T., Black, M.J.: Star: Sparse trained articulated human body regressor. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. pp. 598–613. Springer (2020) 4

45. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: Agora: Avatars in geography optimized for regression analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13468–13478 (2021) 3
46. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10975–10985 (2019) 14
47. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9054–9063 (2021) 4
48. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P.V., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4929–4937 (2016) 3
49. Reddy, N.D., Guigues, L., Pishchulin, L., Eledath, J., Narasimhan, S.G.: Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15190–15200 (2021) 4
50. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610 (2022) 14
51. Saito, S., Yang, J., Ma, Q., Black, M.J.: Scanimate: Weakly supervised learning of skinned clothed avatar networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2886–2897 (2021) 4
52. Shuai, Q., Geng, C., Fang, Q., Peng, S., Shen, W., Zhou, X., Bao, H.: Novel view synthesis of human interactions from sparse multi-view videos. In: ACM SIGGRAPH 2022 Conference Proceedings. SIGGRAPH '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3528233.3530704, https://doi.org/10.1145/3528233.3530704 3, 4, 8, 9
53. Shuai, Q., Yu, Z., Zhou, Z., Fan, L., Yang, H., Yang, C., Zhou, X.: Reconstructing close human interactions from multiple views. ACM Transactions on Graphics (dec 2023). https://doi.org/10.1145/3618336 3, 4
54. Song, J., Chen, X., Hilliges, O.: Human body model fitting by learned gradient descent. In: European Conference on Computer Vision. pp. 744–760. Springer (2020) 14
55. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3d people. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11179–11188 (2021) 3
56. Sun, Y., Liu, W., Bao, Q., Fu, Y., Mei, T., Black, M.J.: Putting people in their place: Monocular regression of 3d people in depth. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13243–13252 (2022) 3
57. Tu, H., Wang, C., Zeng, W.: Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 197–212. Springer (2020) 2, 4, 8, 9
58. Wang, C., Li, J., Liu, W., Qian, C., Lu, C.: Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 242–259. Springer (2020) 3

59. Wang, S., Schwarz, K., Geiger, A., Tang, S.: Arah: Animatable volume rendering of articulated human sdfs. In: European conference on computer vision. pp. 1–19. Springer (2022) 4
60. Wang, T., Zhang, J., Cai, Y., Yan, S., Feng, J.: Direct multi-view multi-person 3d pose estimation. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 13153–13164. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/6da9003b743b65f4c0ccd295cc484e57-Paper.pdf 2, 3, 4, 10, 11
61. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition. pp. 16210–16220 (2022) 4, 8
62. Wu, S., Jin, S., Liu, W., Bai, L., Qian, C., Liu, D., Ouyang, W.: Graph-based 3d multi-person pose estimation using multi-view images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11148–11157 (2021) 2, 3, 4, 10, 11
63. Wu, S., Jin, S., Liu, W., Bai, L., Qian, C., Liu, D., Ouyang, W.: Graph-based 3d multi-person pose estimation using multi-view images. In: ICCV (2021) 10
64. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Ghum & ghuml: Generative 3d human shape and articulated pose models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6184–6193 (2020) 4
65. Ye, H., Zhu, W., Wang, C., Wu, R., Wang, Y.: Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In: European Conference on Computer Vision. pp. 142–159. Springer (2022) 2, 3, 4, 10, 11
66. Yin, Y., Guo, C., Kaufmann, M., Zarate, J.J., Song, J., Hilliges, O.: Hi4d: 4d instance segmentation of close human interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17016–17027 (2023) 5, 9, 11
67. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2148–2157 (2018) 3
68. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11446–11456 (2021) 3
69. Zhang, J., Liu, X., Ye, X., Zhao, F., Zhang, Y., Wu, M., Zhang, Y., Xu, L., Yu, J.: Editable free-viewpoint video using a layered neural representation. ACM Transactions on Graphics (TOG) **40**(4), 1–18 (2021) 3, 4, 6, 7
70. Zhang, Y., Wang, C., Wang, X., Liu, W., Zeng, W.: Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 2613–2626 (2022) 4
71. Zhang, Y., An, L., Yu, T., Li, X., Li, K., Liu, Y.: 4d association graph for realtime multi-person motion capture using multiple video cameras. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1324–1333 (2020) 2, 3, 8, 10, 11
72. Zhang, Y., Li, Z., An, L., Li, M., Yu, T., Liu, Y.: Lightweight multi-person total motion capture using sparse multi-view cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5560–5569 (2021) 4

73. Zheng, Y., Shao, R., Zhang, Y., Yu, T., Zheng, Z., Dai, Q., Liu, Y.: Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6239–6249 (2021) 4, 5
74. Zheng, Z., Zhao, X., Zhang, H., Liu, B., Liu, Y.: Avatarrex: Real-time expressive full-body avatars. arXiv preprint arXiv:2305.04789 (2023) 14
75. Zhou, H., Hong, C., Han, Y., Huang, P., Zhuang, Y.: Mh pose: 3d human pose estimation based on high-quality heatmap. In: 2021 IEEE International Conference on Big Data (Big Data). pp. 3215–3222. IEEE (2021) 4
76. Zhou, Z., Shuai, Q., Wang, Y., Fang, Q., Ji, X., Li, F., Bao, H., Zhou, X.: Quickpose: Real-time multi-view multi-person pose estimation in crowded scenes. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–9 (2022) 3